CHAPTER 1

Introduction

From helping to assess the value of new medical treatments to evaluating the factors that affect our opinions on various controversial issues, scientists today are finding myriad uses for methods of analyzing categorical data. It's primarily for these scientists and their collaborating statisticians – as well as those training to perform these roles – that this book was written. The book provides an introduction to methods for analyzing categorical data. It emphasizes the ideas behind the methods and their interpretations, rather than the theory behind them.

This first chapter reviews the probability distributions most often used for categorical data, such as the *binomial distribution*. It also introduces *maximum likelihood*, the most popular method for estimating parameters. We use this estimate and a related *likelihood function* to conduct statistical inference about proportions. We begin by discussing the major types of categorical data and summarizing the book's outline.

1.1 CATEGORICAL RESPONSE DATA

Let us first define categorical data. A *categorical* variable has a measurement scale consisting of a set of categories. For example, political philosophy may be measured as "liberal," "moderate," or "conservative"; choice of accommodation might use categories "house," "condominium," "apartment"; a diagnostic test to detect e-mail spam might classify an incoming e-mail message as "spam" or "legitimate e-mail."

Categorical scales are pervasive in the social sciences for measuring attitudes and opinions. Categorical scales also occur frequently in the health sciences, for measuring responses such as whether a patient survives an operation (yes, no), severity of an injury (none, mild, moderate, severe), and stage of a disease (initial, advanced).

Although categorical variables are common in the social and health sciences, they are by no means restricted to those areas. They frequently occur in the behavioral

An Introduction to Categorical Data Analysis, Second Edition. By Alan Agresti

sciences (e.g., categories "schizophrenia," "depression," "neurosis" for diagnosis of type of mental illness), public health (e.g., categories "yes" and "no" for whether awareness of AIDS has led to increased use of condoms), zoology (e.g., categories "fish," "invertebrate," "reptile" for alligators' primary food choice), education (e.g., categories "correct" and "incorrect" for students' responses to an exam question), and marketing (e.g., categories "Brand A," "Brand B," and "Brand C" for consumers' preference among three leading brands of a product). They even occur in highly quantitative fields such as engineering sciences and industrial quality control, when items are classified according to whether or not they conform to certain standards.

1.1.1 Response/Explanatory Variable Distinction

Most statistical analyses distinguish between *response* variables and *explanatory* variables. For instance, regression models describe how the distribution of a continuous response variable, such as annual income, changes according to levels of explanatory variables, such as number of years of education and number of years of job experience. The response variable is sometimes called the *dependent variable* or *Y variable*, and the explanatory variable is sometimes called the *independent variable* or *X variable*.

The subject of this text is the analysis of categorical response variables. The categorical variables listed in the previous subsection are response variables. In some studies, they might also serve as explanatory variables. Statistical models for categorical response variables analyze how such responses are influenced by explanatory variables. For example, a model for political philosophy could use predictors such as annual income, attained education, religious affiliation, age, gender, and race. The explanatory variables can be categorical or continuous.

1.1.2 Nominal/Ordinal Scale Distinction

Categorical variables have two main types of measurement scales. Many categorical scales have a natural ordering. Examples are attitude toward legalization of abortion (disapprove in all cases, approve only in certain cases, approve in all cases), appraisal of a company's inventory level (too low, about right, too high), response to a medical treatment (excellent, good, fair, poor), and frequency of feeling symptoms of anxiety (never, occasionally, often, always). Categorical variables having ordered scales are called *ordinal* variables.

Categorical variables having unordered scales are called *nominal* variables. Examples are religious affiliation (categories Catholic, Jewish, Protestant, Muslim, other), primary mode of transportation to work (automobile, bicycle, bus, subway, walk), favorite type of music (classical, country, folk, jazz, rock), and favorite place to shop (local mall, local downtown, Internet, other).

For nominal variables, the order of listing the categories is irrelevant. The statistical analysis should not depend on that ordering. Methods designed for nominal variables give the same results no matter how the categories are listed. Methods designed for

ordinal variables utilize the category ordering. Whether we list the categories from low to high or from high to low is irrelevant in terms of substantive conclusions, but results of ordinal analyses would change if the categories were reordered in any other way.

Methods designed for ordinal variables *cannot* be used with nominal variables, since nominal variables do not have ordered categories. Methods designed for nominal variables *can* be used with nominal or ordinal variables, since they only require a categorical scale. When used with ordinal variables, however, they do not use the information about that ordering. This can result in serious loss of power. It is usually best to apply methods appropriate for the actual scale.

Categorical variables are often referred to as *qualitative*, to distinguish them from numerical-valued or *quantitative* variables such as weight, age, income, and number of children in a family. However, we will see it is often advantageous to treat ordinal data in a quantitative manner, for instance by assigning ordered scores to the categories.

1.1.3 Organization of this Book

Chapters 1 and 2 describe some standard methods of categorical data analysis developed prior to about 1960. These include basic analyses of association between two categorical variables.

Chapters 3–7 introduce models for categorical responses. These models resemble regression models for continuous response variables. In fact, Chapter 3 shows they are special cases of a generalized class of linear models that also contains the usual normal-distribution-based regression models. The main emphasis in this book is on *logistic regression* models. Applying to response variables that have two outcome categories, they are the focus of Chapters 4 and 5. Chapter 6 presents extensions to multicategory responses, both nominal and ordinal. Chapter 7 introduces *loglinear* models, which analyze associations among multiple categorical response variables.

The methods in Chapters 1–7 assume that observations are independent. Chapters 8–10 discuss logistic regression models that apply when some observations are correlated, such as with repeated measurement of subjects in longitudinal studies. An important special case is matched pairs that result from observing a categorical response for the same subjects at two separate times. The book concludes (Chapter 11) with a historical overview of categorical data methods.

Most methods for categorical data analysis require extensive computations. The Appendix discusses the use of SAS statistical software. A companion website for the book, http://www.stat.ufl.edu/∼aa/intro-cda/software.html, discusses other software.

1.2 PROBABILITY DISTRIBUTIONS FOR CATEGORICAL DATA

Inferential statistical analyses require assumptions about the probability distribution of the response variable. For regression and analysis of variance (ANOVA)

models for continuous data, the normal distribution plays a central role. This section presents the key distributions for categorical data: the *binomial* and *multinomial* distributions.

1.2.1 Binomial Distribution

Often, categorical data result from n independent and identical trials with two possible outcomes for each, referred to as "success" and "failure." These are generic labels, and the "success" outcome need not be a preferred result. *Identical trials* means that the probability of success is the same for each trial. *Independent trials* means the response outcomes are independent random variables. In particular, the outcome of one trial does not affect the outcome of another. These are often called *Bernoulli trials*. Let π denote the probability of success for a given trial. Let Y denote the number of successes out of the n trials.

Under the assumption of n independent, identical trials, Y has the *binomial distribution* with index n and parameter π. You are probably familiar with this distribution, but we review it briefly here. The probability of outcome y for Y equals

$$P(y) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y}, \quad y = 0, 1, 2, \ldots, n \tag{1.1}$$

To illustrate, suppose a quiz has 10 multiple-choice questions, with five possible answers for each. A student who is completely unprepared randomly guesses the answer for each question. Let Y denote the number of correct responses. The probability of a correct response is 0.20 for a given question, so $n = 10$ and $\pi = 0.20$. The probability of $y = 0$ correct responses, and hence $n - y = 10$ incorrect ones, equals

$$P(0) = [10!/(0!10!)](0.20)^0(0.80)^{10} = (0.80)^{10} = 0.107.$$

The probability of 1 correct response equals

$$P(1) = [10!/(1!9!)](0.20)^1(0.80)^9 = 10(0.20)(0.80)^9 = 0.268.$$

Table 1.1 shows the entire distribution. For contrast, it also shows the distributions when $\pi = 0.50$ and when $\pi = 0.80$.

The binomial distribution for n trials with parameter π has mean and standard deviation

$$E(Y) = \mu = n\pi, \quad \sigma = \sqrt{n\pi(1-\pi)}$$

The binomial distribution in Table 1.1 has $\mu = 10(0.20) = 2.0$ and $\sigma = \sqrt{[10(0.20)(0.80)]} = 1.26$.

The binomial distribution is always symmetric when $\pi = 0.50$. For fixed n, it becomes more skewed as π moves toward 0 or 1. For fixed π, it becomes more

Table 1.1. Binomial Distribution with $n = 10$ and $\pi = 0.20, 0.50$, and 0.80. The Distribution is Symmetric when $\pi = 0.50$

y	$P(y)$ when $\pi = 0.20$	$P(y)$ when $\pi = 0.50$	$P(y)$ when $\pi = 0.80$
0	0.107	0.001	0.000
1	0.268	0.010	0.000
2	0.302	0.044	0.000
3	0.201	0.117	0.001
4	0.088	0.205	0.005
5	0.027	0.246	0.027
6	0.005	0.205	0.088
7	0.001	0.117	0.201
8	0.000	0.044	0.302
9	0.000	0.010	0.268
10	0.000	0.001	0.107

bell-shaped as n increases. When n is large, it can be approximated by a normal distribution with $\mu = n\pi$ and $\sigma = \sqrt{[n\pi(1-\pi)]}$. A guideline is that the expected number of outcomes of the two types, $n\pi$ and $n(1-\pi)$, should both be at least about 5. For $\pi = 0.50$ this requires only $n \geq 10$, whereas $\pi = 0.10$ (or $\pi = 0.90$) requires $n \geq 50$. When π gets nearer to 0 or 1, larger samples are needed before a symmetric, bell shape occurs.

1.2.2 Multinomial Distribution

Some trials have more than two possible outcomes. For example, the outcome for a driver in an auto accident might be recorded using the categories "uninjured," "injury not requiring hospitalization," "injury requiring hospitalization," "fatality." When the trials are independent with the same category probabilities for each trial, the distribution of counts in the various categories is the *multinomial*.

Let c denote the number of outcome categories. We denote their probabilities by $\{\pi_1, \pi_2, \ldots, \pi_c\}$, where $\sum_j \pi_j = 1$. For n independent observations, the multinomial probability that n_1 fall in category 1, n_2 fall in category 2, $\ldots$, n_c fall in category c, where $\sum_j n_j = n$, equals

$$P(n_1, n_2, \ldots, n_c) = \left(\frac{n!}{n_1!n_2!\cdots n_c!}\right)\pi_1^{n_1}\pi_2^{n_2}\cdots\pi_c^{n_c}$$

The binomial distribution is the special case with $c = 2$ categories. We will not need to use this formula, as we will focus instead on sampling distributions of useful statistics computed from data assumed to have the multinomial distribution. We present it here merely to show how the binomial formula generalizes to several outcome categories.

The multinomial is a multivariate distribution. The marginal distribution of the count in any particular category is binomial. For category j, the count n_j has mean $n\pi_j$ and standard deviation $\sqrt{[n\pi_j(1-\pi_j)]}$. Most methods for categorical data assume the binomial distribution for a count in a single category and the multinomial distribution for a set of counts in several categories.

1.3 STATISTICAL INFERENCE FOR A PROPORTION

In practice, the parameter values for the binomial and multinomial distributions are unknown. Using sample data, we estimate the parameters. This section introduces the estimation method used in this text, called *maximum likelihood*. We illustrate this method for the binomial parameter.

1.3.1 Likelihood Function and Maximum Likelihood Estimation

The parametric approach to statistical modeling assumes a family of probability distributions, such as the binomial, for the response variable. For a particular family, we can substitute the observed data into the formula for the probability function and then view how that probability depends on the unknown parameter value. For example, in $n = 10$ trials, suppose a binomial count equals $y = 0$. From the binomial formula (1.1) with parameter π, the probability of this outcome equals

$$P(0) = [10!/(0!)(10!)]\pi^0(1-\pi)^{10} = (1-\pi)^{10}$$

This probability is defined for all the potential values of π between 0 and 1.

The probability of the observed data, expressed as a function of the parameter, is called the *likelihood function*. With $y = 0$ successes in $n = 10$ trials, the binomial likelihood function is $\ell(\pi) = (1-\pi)^{10}$. It is defined for π between 0 and 1. From the likelihood function, if $\pi = 0.40$ for instance, the probability that $Y = 0$ is $\ell(0.40) = (1-0.40)^{10} = 0.006$. Likewise, if $\pi = 0.20$ then $\ell(0.20) = (1-0.20)^{10} = 0.107$, and if $\pi = 0.0$ then $\ell(0.0) = (1-0.0)^{10} = 1.0$. Figure 1.1 plots this likelihood function.

The *maximum likelihood estimate* of a parameter is the parameter value for which the probability of the observed data takes its greatest value. It is the parameter value at which the likelihood function takes its maximum. Figure 1.1 shows that the likelihood function $\ell(\pi) = (1-\pi)^{10}$ has its maximum at $\pi = 0.0$. Thus, when $n = 10$ trials have $y = 0$ successes, the maximum likelihood estimate of π equals 0.0. This means that the result $y = 0$ in $n = 10$ trials is more likely to occur when $\pi = 0.00$ than when π equals any other value.

In general, for the binomial outcome of y successes in n trials, the maximum likelihood estimate of π equals $p = y/n$. This is the sample proportion of successes for the n trials. If we observe $y = 6$ successes in $n = 10$ trials, then the maximum likelihood estimate of π equals $p = 6/10 = 0.60$. Figure 1.1 also plots the

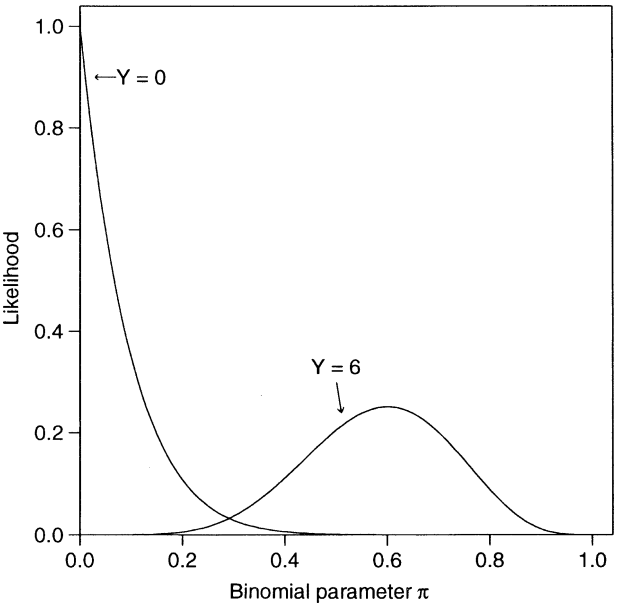

Figure 1.1. Binomial likelihood functions for $y = 0$ successes and for $y = 6$ successes in $n = 10$ trials.

likelihood function when $n = 10$ with $y = 6$, which from formula (1.1) equals $\ell(\pi) = [10!/(6!)(4!)]\pi^6(1-\pi)^4$. The maximum value occurs when $\pi = 0.60$. The result $y = 6$ in $n = 10$ trials is more likely to occur when $\pi = 0.60$ than when π equals any other value.

Denote each success by a 1 and each failure by a 0. Then the sample proportion equals the sample mean of the results of the individual trials. For instance, for four failures followed by six successes in 10 trials, the data are 0,0,0,0,1,1,1,1,1,1, and the sample mean is

$$p = (0+0+0+0+1+1+1+1+1+1)/10 = 0.60.$$

Thus, results that apply to sample means with random sampling, such as the Central Limit Theorem (large-sample normality of its sampling distribution) and the Law of Large Numbers (convergence to the population mean as n increases) apply also to sample proportions.

The abbreviation *ML* symbolizes the term *maximum likelihood*. The ML estimate is often denoted by the parameter symbol with a ^(a "hat") over it. The ML estimate of the binomial parameter π, for instance, is often denoted by $\hat{\pi}$, called *pi-hat*.

Before we observe the data, the value of the ML estimate is unknown. The estimate is then a variate having some sampling distribution. We refer to this variate as an *estimator* and its value for observed data as an *estimate*. Estimators based on the method of maximum likelihood are popular because they have good large-sample behavior. Most importantly, it is not possible to find good estimators that are more

precise, in terms of having smaller large-sample standard errors. Also, large-sample distributions of ML estimators are usually approximately normal. The estimators reported in this text use this method.

1.3.2 Significance Test About a Binomial Proportion

For the binomial distribution, we now use the ML estimator in statistical inference for the parameter π. The ML estimator is the sample proportion, p. The sampling distribution of the sample proportion p has mean and standard error

$$E(p) = \pi, \quad \sigma(p) = \sqrt{\frac{\pi(1-\pi)}{n}}$$

As the number of trials n increases, the standard error of p decreases toward zero; that is, the sample proportion tends to be closer to the parameter value π. The sampling distribution of p is approximately normal for large n. This suggests large-sample inferential methods for π.

Consider the null hypothesis H_0: $\pi = \pi_0$ that the parameter equals some fixed value, π_0. The test statistic

$$z = \frac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} \tag{1.2}$$

divides the difference between the sample proportion p and the null hypothesis value π_0 by the *null standard error* of p. The null standard error is the one that holds under the assumption that the null hypothesis is true. For large samples, the null sampling distribution of the z test statistic is the standard normal – the normal distribution having a mean of 0 and standard deviation of 1. The z test statistic measures the number of standard errors that the sample proportion falls from the null hypothesized proportion.

1.3.3 Example: Survey Results on Legalizing Abortion

Do a majority, or minority, of adults in the United States believe that a pregnant woman should be able to obtain an abortion? Let π denote the proportion of the American adult population that responds "yes" to the question, "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if she is married and does not want any more children." We test H_0: $\pi = 0.50$ against the two-sided alternative hypothesis, H_a: $\pi \neq 0.50$.

This item was one of many included in the 2002 General Social Survey. This survey, conducted every other year by the National Opinion Research Center (NORC) at the University of Chicago, asks a sample of adult American subjects their opinions about a wide variety of issues. (It is a multistage sample, but has characteristics

similar to a simple random sample.) You can view responses to surveys since 1972 at http://sda.berkeley.edu/GSS. Of 893 respondents to this question in 2002, 400 replied "yes" and 493 replied "no".

The sample proportion of "yes" responses was $p = 400/893 = 0.448$. For a sample of size $n = 893$, the null standard error of p equals $\sqrt{[(0.50)(0.50)/893]} = 0.0167$. The test statistic is

$$z = (0.448 - 0.50)/0.0167 = -3.1$$

The two-sided P-value is the probability that the absolute value of a standard normal variate exceeds 3.1, which is $P = 0.002$. There is strong evidence that, in 2002, $\pi < 0.50$, that is, that fewer than half of Americans favored legal abortion in this situation. In some other situations, such as when the mother's health was endangered, an overwhelming majority favored legalized abortion. Responses depended strongly on the question wording.

1.3.4 Confidence Intervals for a Binomial Proportion

A significance test merely indicates whether a particular value for a parameter (such as 0.50) is plausible. We learn more by constructing a confidence interval to determine the range of plausible values. Let SE denote the estimated standard error of p. A large-sample $100(1 - \alpha)\%$ confidence interval for π has the formula

$$p \pm z_{\alpha/2}(SE), \quad \text{with } SE = \sqrt{p(1-p)/n} \tag{1.3}$$

where $z_{\alpha/2}$ denotes the standard normal percentile having right-tail probability equal to $\alpha/2$; for example, for 95% confidence, $\alpha = 0.05$, $z_{\alpha/2} = z_{0.025} = 1.96$. This formula substitutes the sample proportion p for the unknown parameter π in $\sigma(p) = \sqrt{[\pi(1-\pi)/n]}$.

For the attitudes about abortion example just discussed, $p = 0.448$ for $n = 893$ observations. The 95% confidence interval equals

$$0.448 \pm 1.96\sqrt{(0.448)(0.552)/893}, \quad \text{which is } 0.448 \pm 0.033, \quad \text{or } (0.415, 0.481)$$

We can be 95% confident that the population proportion of Americans in 2002 who favored legalized abortion for married pregnant women who do not want more children is between 0.415 and 0.481.

Formula (1.3) is simple. Unless π is close to 0.50, however, it does not work well unless n is very large. Consider its actual coverage probability, that is, the probability that the method produces an interval that captures the true parameter value. This may be quite a bit less than the nominal value (such as 95%). It is especially poor when π is near 0 or 1.

A better way to construct confidence intervals uses a duality with significance tests. This confidence interval consists of all values π_0 for the null hypothesis parameter

that are judged plausible in the z test of the previous subsection. A 95% confidence interval contains all values π_0 for which the two-sided P-value exceeds 0.05. That is, it contains all values that are "not rejected" at the 0.05 significance level. These are the null values that have test statistic z less than 1.96 in absolute value. This alternative method does not require estimation of π in the standard error, since the standard error in the test statistic uses the null value π_0.

To illustrate, suppose a clinical trial to evaluate a new treatment has nine successes in the first 10 trials. For a sample proportion of $p = 0.90$ based on $n = 10$, the value $\pi_0 = 0.596$ for the null hypothesis parameter leads to the test statistic value

$$z = (0.90 - 0.596)/\sqrt{(0.596)(0.404)/10} = 1.96$$

and a two-sided P-value of $P = 0.05$. The value $\pi_0 = 0.982$ leads to

$$z = (0.90 - 0.982)/\sqrt{(0.982)(0.018)/100} = -1.96$$

and also a two-sided P-value of $P = 0.05$. (We explain in the following paragraph how to find 0.596 and 0.982.) All π_0 values between 0.596 and 0.982 have $|z| < 1.96$ and $P > 0.05$. So, the 95% confidence interval for π equals (0.596, 0.982). By contrast, the method (1.3) using the *estimated* standard error gives confidence interval $0.90 \pm 1.96\sqrt{[(0.90)(0.10)/10]}$, which is (0.714, 1.086). However, it works poorly to use the sample proportion as the midpoint of the confidence interval when the parameter may fall near the boundary values of 0 or 1.

For given p and n, the π_0 values that have test statistic value $z = \pm 1.96$ are the solutions to the equation

$$\frac{|\ p - \pi_0\ |}{\sqrt{\pi_0(1 - \pi_0)/n}} = 1.96$$

for π_0. To solve this for π_0, squaring both sides gives an equation that is quadratic in π_0 (see Exercise 1.18). The results are available with some software, such as an *R* function available at http://www.stat.ufl.edu/~aa/cda/software.html.

Here is a simple alternative interval that approximates this one, having a similar midpoint in the 95% case but being a bit wider: Add 2 to the number of successes and 2 to the number of failures (and thus 4 to n) and then use the ordinary formula (1.3) with the estimated standard error. For example, with nine successes in 10 trials, you find $p = (9 + 2)/(10 + 4) = 0.786$, $SE = \sqrt{[0.786(0.214)/14]} = 0.110$, and obtain confidence interval (0.57, 1.00). This simple method, sometimes called the *Agresti–Coull confidence interval*, works well even for small samples.[1]

[1] A. Agresti and B. Coull, *Am. Statist.*, **52**: 119–126, 1998.

1.4 MORE ON STATISTICAL INFERENCE FOR DISCRETE DATA

We have just seen how to construct a confidence interval for a proportion using an estimated standard error or by inverting results of a significance test using the null standard error. In fact, there are three ways of using the likelihood function to conduct inference (confidence intervals and significance tests) about parameters. We finish the chapter by summarizing these methods. They apply to any parameter in a statistical model, but we will illustrate using the binomial parameter.

1.4.1 Wald, Likelihood-Ratio, and Score Inference

Let β denote an arbitrary parameter. Consider a significance test of H_0: $\beta = \beta_0$ (such as H_0: $\beta = 0$, for which $\beta_0 = 0$).

The simplest test statistic uses the large-sample normality of the ML estimator $\hat{\beta}$. Let SE denote the standard error of $\hat{\beta}$, evaluated by substituting the ML estimate for the unknown parameter in the expression for the true standard error. (For example, for the binomial parameter π, $SE = \sqrt{[p(1-p)/n]}$.) When H_0 is true, the test statistic

$$z = (\hat{\beta} - \beta_0)/SE$$

has approximately a standard normal distribution. Equivalently, z^2 has approximately a chi-squared distribution with $df = 1$. This type of statistic, which uses the standard error evaluated at the ML estimate, is called a *Wald statistic*. The z or chi-squared test using this test statistic is called a *Wald test*.

You can refer z to the standard normal table to get one-sided or two-sided P-values. Equivalently, for the two-sided alternative H_0: $\beta \neq \beta_0$, z^2 has a chi-squared distribution with $df = 1$. The P-value is then the right-tail chi-squared probability above the observed value. The two-tail probability beyond $\pm z$ for the standard normal distribution equals the right-tail probability above z^2 for the chi-squared distribution with $df = 1$. For example, the two-tail standard normal probability of 0.05 that falls below -1.96 and above 1.96 equals the right-tail chi-squared probability above $(1.96)^2 = 3.84$ when $df = 1$.

An alternative test uses the likelihood function through the ratio of two maximizations of it: (1) the maximum over the possible parameter values that assume the null hypothesis, (2) the maximum over the larger set of possible parameter values, permitting the null or the alternative hypothesis to be true. Let ℓ_0 denote the maximized value of the likelihood function under the null hypothesis, and let ℓ_1 denote the maximized value more generally. For instance, when there is a single parameter β, ℓ_0 is the likelihood function calculated at β_0, and ℓ_1 is the likelihood function calculated at the ML estimate $\hat{\beta}$. Then ℓ_1 is always at least as large as ℓ_0, because ℓ_1 refers to maximizing over a larger set of possible parameter values.

The *likelihood-ratio* test statistic equals

$$-2\log(\ell_0/\ell_1)$$

In this text, we use the *natural log*, often abbreviated on calculators by LN. If the maximized likelihood is much larger when the parameters are not forced to satisfy H_0, then the ratio ℓ_0/ℓ_1 is far below 1. The test statistic $-2\log(\ell_0/\ell_1)$ must be nonnegative, and relatively small values of ℓ_0/ℓ_1 yield large values of $-2\log(\ell_0/\ell_1)$ and strong evidence against H_0. The reason for taking the log transform and doubling is that it yields an approximate chi-squared sampling distribution. Under H_0: $\beta = \beta_0$, the likelihood-ratio test statistic has a large-sample chi-squared distribution with $df = 1$. Software can find the maximized likelihood values and the likelihood-ratio test statistic.

A third possible test is called the *score test*. We will not discuss the details except to say that it finds standard errors under the assumption that the null hypothesis holds. For example, the z test (1.2) for a binomial parameter that uses the standard error $\sqrt{[\pi_0(1-\pi_0)/n]}$ is a score test.

The Wald, likelihood-ratio, and score tests are the three major ways of constructing significance tests for parameters in statistical models. For ordinary regression models assuming a normal distribution for Y, the three tests provide identical results. In other cases, for large samples they have similar behavior when H_0 is true.

When you use any of these tests, the P-value that you find or software reports is an approximation for the true P-value. This is because the normal (or chi-squared) sampling distribution used is a large-sample approximation for the true sampling distribution. Thus, when you report a P-value, it is overly optimistic to use many decimal places. If you are lucky, the P-value approximation is good to the second decimal place. So, for a P-value that software reports as 0.028374, it makes more sense to report it as 0.03 (or, at best, 0.028) rather than 0.028374. An exception is when the P-value is zero to many decimal places, in which case it is sensible to report it as $P < 0.001$ or $P < 0.0001$. In any case, a P-value merely summarizes the strength of evidence against the null hypothesis, and accuracy to two or three decimal places is sufficient for this purpose.

Each method has a corresponding confidence interval. This is based on inverting results of the significance test: The 95% confidence interval for a parameter β is the set of β_0 values for the significance test of H_0: $\beta = \beta_0$ such that the P-value is larger than 0.05. For example, the 95% *Wald confidence interval* is the set of β_0 values for which $z = (\hat{\beta} - \beta_0)/SE$ has $|z| < 1.96$. It equals $\hat{\beta} \pm 1.96(SE)$. For a binomial proportion, the score confidence interval is the one discussed in Section 1.3.4 that has endpoints that are π_0 values having P-value 0.05 in the z-test using the null standard error.

1.4.2 Wald, Score, and Likelihood-Ratio Inference for Binomial Parameter

We illustrate the Wald, likelihood-ratio, and score tests by testing H_0: $\pi = 0.50$ against H_a: $\pi \neq 0.50$ for the example mentioned near the end of Section 1.3.4 of a clinical trial that has nine successes in the first 10 trials. The sample proportion is $p = 0.90$ based on $n = 10$.

For the Wald test of H_0: $\pi = 0.50$, the estimated standard error is $SE = \sqrt{[0.90(0.10)/10]} = 0.095$. The z test statistic is

$$z = (0.90 - 0.50)/0.095 = 4.22$$

The corresponding chi-squared statistic is $(4.22)^2 = 17.8$ ($df = 1$). The P-value <0.001.

For the score test of H_0: $\pi = 0.50$, the null standard error is $\sqrt{[0.50(0.50)/10]} = 0.158$. The z test statistic is

$$z = (0.90 - 0.50)/0.158 = 2.53$$

The corresponding chi-squared statistic is $(2.53)^2 = 6.4$ ($df = 1$). The P-value $= 0.011$.

Finally, consider the likelihood-ratio test. When H_0: $\pi = 0.50$ is true, the binomial probability of the observed result of nine successes is $\ell_0 = [10!/9!1!](0.50)^9(0.50)^1 = 0.00977$. The likelihood-ratio test compares this to the value of the likelihood function at the ML estimate of $p = 0.90$, which is $\ell_1 = [10!/9!1!](0.90)^9(0.10)^1 = 0.3874$. The likelihood-ratio test statistic equals

$$-2\log(\ell_0/\ell_1) = -2\log(0.00977/0.3874) = -2\log(0.0252) = 7.36$$

From the chi-squared distribution with $df = 1$, this statistic has P-value $= 0.007$.

When the sample size is small to moderate, the Wald test is the least reliable of the three tests. We should not trust it for such a small n as in this example ($n = 10$). Likelihood-ratio inference and score-test based inference are better in terms of actual error probabilities coming close to matching nominal levels. A marked divergence in the values of the three statistics indicates that the distribution of the ML estimator may be far from normality. In that case, small-sample methods are more appropriate than large-sample methods.

1.4.3 Small-Sample Binomial Inference

For inference about a proportion, the large-sample two-sided z score test and the confidence interval based on that test (using the null hypothesis standard error) perform reasonably well when $n\pi \geq 5$ and $n(1 - \pi) \geq 5$. When π_0 is not near 0.50 the normal P-value approximation is better for the test with a two-sided alternative than for a one-sided alternative; a probability that is "too small" in one tail tends to be approximately counter-balanced by a probability that is "too large" in the other tail.

For small sample sizes, it is safer to use the binomial distribution directly (rather than a normal approximation) to calculate P-values. To illustrate, consider testing H_0: $\pi = 0.50$ against H_a: $\pi > 0.50$ for the example of a clinical trial to evaluate a new treatment, when the number of successes $y = 9$ in $n = 10$ trials. The exact P-value, based on the right tail of the null binomial distribution with $\pi = 0.50$, is

$$P(Y \geq 9) = [10!/9!1!](0.50)^9(0.50)^1 + [10!/10!0!](0.50)^{10}(0.50)^0 = 0.011$$

For the two sided alternative H_a: $\pi \neq 0.50$, the P-value is

$$P(Y \geq 9 \text{ or } Y \leq 1) = 2 \times P(Y \geq 9) = 0.021$$

1.4.4 Small-Sample Discrete Inference is Conservative*

Unfortunately, with discrete probability distributions, small-sample inference using the ordinary P-value is *conservative*. This means that when H_0 is true, the P-value is ≤ 0.05 (thus leading to rejection of H_0 at the 0.05 significance level) not *exactly* 5% of the time, but typically *less* than 5% of the time. Because of the discreteness, it is usually not possible for a P-value to achieve the desired significance level exactly. Then, the actual probability of type I error is less than 0.05.

For example, consider testing H_0: $\pi = 0.50$ against H_a: $\pi > 0.50$ for the clinical trial example with $y = 9$ successes in $n = 10$ trials. Table 1.1 showed the binomial distribution with $n = 10$ and $\pi = 0.50$. Table 1.2 shows it again with the corresponding P-values (right-tail probabilities) for this one-sided alternative. The P-value is ≤ 0.05 when $y = 9$ or 10. This happens with probability $0.010 + 0.001 = 0.011$. Thus, the probability of getting a P-value ≤ 0.05 is only 0.011. For a desired significance level of 0.05, the actual probability of type I error is 0.011. The actual probability of type I error is much smaller than the intended one.

This illustrates an awkward aspect of significance testing when the test statistic has a discrete distribution. For test statistics having a *continuous* distribution, the P-value has a *uniform* null distribution over the interval [0, 1]. That is, when H_0 is true, the P-value is equally likely to fall anywhere between 0 and 1. Then, the probability that the P-value falls below 0.05 equals exactly 0.05, and the expected value of the P-value is exactly 0.50. For a test statistic having a *discrete* distribution, the null distribution of the P-value is discrete and has an expected value greater than 0.50.

For example, for the one-sided test summarized above, the P-value equals 1.000 with probability $P(0) = 0.001$, it equals 0.999 with probability $P(1) = 0.010, \ldots,$ and it equals 0.001 with probability $P(10) = 0.001$. From the table, the null expected

Table 1.2. Null Binomial Distribution and One-Sided P-values for Testing H_0: $\pi = 0.50$ against H_a: $\pi > 0.50$ with $n = 10$

y	$P(y)$	P-value	Mid P-value
0	0.001	1.000	0.9995
1	0.010	0.999	0.994
2	0.044	0.989	0.967
3	0.117	0.945	0.887
4	0.205	0.828	0.726
5	0.246	0.623	0.500
6	0.205	0.377	0.274
7	0.117	0.172	0.113
8	0.044	0.055	0.033
9	0.010	0.011	0.006
10	0.001	0.001	0.0005

value of the P-value is

$$\sum P \times \text{Prob}(P) = 1.000(0.001) + 0.999(0.010) + \cdots + 0.001(0.001) = 0.59$$

In this average sense, P-values for discrete distributions tend to be too large.

1.4.5 Inference Based on the Mid P-value*

With small samples of discrete data, many statisticians prefer to use a different type of P-value. Called the *mid P-value*, it adds only *half* the probability of the observed result to the probability of the more extreme results. To illustrate, in the above example with $y = 9$ successes in $n = 10$ trials, the ordinary P-value for H_a: $\pi > 0.50$ is $P(9) + P(10) = 0.010 + 0.001 = 0.011$. The mid P-value is $P(9)/2 + P(10) = 0.010/2 + 0.001 = 0.006$. Table 1.2 also shows the mid P-values for the possible y values when $n = 10$.

Tests using the mid P-value are, on the average, less conservative than tests using the ordinary P-value. The mid P-value has a null expected value of 0.50, the same as the regular P-value for continuous variates. Also, the two separate one-sided mid P-values sum to 1.0. For example, for $y = 9$ when $n = 10$, for H_a: $\pi > 0.50$ the ordinary P-value is

$$\text{right-tail } P\text{-value} = P(9) + P(10) = 0.011$$

and for H_a: $\pi < 0.50$ it is

$$\text{left-tail } P\text{-value} = P(0) + P(1) + \cdots + P(9) = 0.999$$

That is, $P(9)$ gets counted in each tail for each P-value. By contrast, for H_a: $\pi > 0.50$, the mid P-value is

$$\text{right-tail mid } P\text{-value} = P(9)/2 + P(10) = 0.006$$

and for H_a: $\pi < 0.50$ it is

$$\text{left-tail mid } P\text{-value} = P(0) + P(1) + \cdots + P(9)/2 = 0.994$$

and these one-sided mid P-values sum to 1.0.

The two-sided P-value for the large-sample z score test approximates the two-sided mid P-value in the small-sample binomial test. For example, with $y = 9$ in $n = 10$ trials for H_0: $\pi = 0.50$, $z = (0.90 - 0.50)/\sqrt{[0.50(0.50)/10]} = 2.53$ has two-sided P-value $= 0.0114$. The two-sided mid P-value is $2[P(9)/2 + P(10)] = 0.0117$.

For small samples, one can construct confidence intervals by inverting results of significance tests that use the binomial distribution, rather than a normal approximation. Such inferences are very conservative when the test uses the ordinary P-value. We recommend inverting instead the binomial test using the mid P-value. The mid-P confidence interval is the set of π_0 values for a two-sided test in which the mid P-value using the binomial distribution exceeds 0.05. This is available in some software, such as an *R* function (written by A. Gottard) at http://www.stat.ufl.edu/~aa/cda/software.html.

1.4.6 Summary

This chapter has introduced the key distributions for categorical data analysis: the binomial and the multinomial. It has also introduced maximum likelihood estimation and illustrated its use for proportion data using Wald, likelihood-ratio, and score methods of inference. The rest of the text uses ML inference for binomial and multinomial parameters in a wide variety of contexts.

PROBLEMS

1.1 In the following examples, identify the response variable and the explanatory variables.

- **a.** Attitude toward gun control (favor, oppose), Gender (female, male), Mother's education (high school, college).
- **b.** Heart disease (yes, no), Blood pressure, Cholesterol level.
- **c.** Race (white, nonwhite), Religion (Catholic, Jewish, Protestant), Vote for president (Democrat, Republican, Other), Annual income.
- **d.** Marital status (married, single, divorced, widowed), Quality of life (excellent, good, fair, poor).

1.2 Which scale of measurement is most appropriate for the following variables – nominal, or ordinal?

- **a.** Political party affiliation (Democrat, Republican, unaffiliated).
- **b.** Highest degree obtained (none, high school, bachelor's, master's, doctorate).
- **c.** Patient condition (good, fair, serious, critical).
- **d.** Hospital location (London, Boston, Madison, Rochester, Toronto).
- **e.** Favorite beverage (beer, juice, milk, soft drink, wine, other).
- **f.** How often feel depressed (never, occasionally, often, always).

1.3 Each of 100 multiple-choice questions on an exam has four possible answers but one correct response. For each question, a student randomly selects one response as the answer.

a. Specify the distribution of the student's number of correct answers on the exam.

b. Based on the mean and standard deviation of that distribution, would it be surprising if the student made at least 50 correct responses? Explain your reasoning.

1.4 A coin is flipped twice. Let $Y =$ number of heads obtained, when the probability of a head for a flip equals π.

a. Assuming $\pi = 0.50$, specify the probabilities for the possible values for Y, and find the distribution's mean and standard deviation.

b. Find the binomial probabilities for Y when π equals (i) 0.60, (ii) 0.40.

c. Suppose you observe $y = 1$ and do not know π. Calculate and sketch the likelihood function.

d. Using the plotted likelihood function from (c), show that the ML estimate of π equals 0.50.

1.5 Refer to the previous exercise. Suppose $y = 0$ in 2 flips. Find the ML estimate of π. Does this estimate seem "reasonable"? Why? [The *Bayesian* estimator is an alternative one that combines the sample data with your prior beliefs about the parameter value. It provides a nonzero estimate of π, equaling $(y + 1)/(n + 2)$ when your prior belief is that π is equally likely to be anywhere between 0 and 1.]

1.6 Genotypes AA, Aa, and aa occur with probabilities (π_1, π_2, π_3). For $n = 3$ independent observations, the observed frequencies are (n_1, n_2, n_3).

a. Explain how you can determine n_3 from knowing n_1 and n_2. Thus, the multinomial distribution of (n_1, n_2, n_3) is actually two-dimensional.

b. Show the set of all possible observations, (n_1, n_2, n_3) with $n = 3$.

c. Suppose $(\pi_1, \pi_2, \pi_3) = (0.25, 0.50, 0.25)$. Find the multinomial probability that $(n_1, n_2, n_3) = (1, 2, 0)$.

d. Refer to (c). What probability distribution does n_1 alone have? Specify the values of the sample size index and parameter for that distribution.

1.7 In his autobiography *A Sort of Life*, British author Graham Greene described a period of severe mental depression during which he played Russian Roulette. This "game" consists of putting a bullet in one of the six chambers of a pistol, spinning the chambers to select one at random, and then firing the pistol once at one's head.

a. Greene played this game six times, and was lucky that none of them resulted in a bullet firing. Find the probability of this outcome.

b. Suppose one kept playing this game until the bullet fires. Let Y denote the number of the game on which the bullet fires. Argue that the probability of

the outcome y equals $(5/6)^{y-1}(1/6)$, for $y = 1, 2, 3, \ldots$. (This is called the *geometric distribution*.)

1.8 When the 2000 General Social Survey asked subjects whether they would be willing to accept cuts in their standard of living to protect the environment, 344 of 1170 subjects said "yes."

a. Estimate the population proportion who would say "yes."

b. Conduct a significance test to determine whether a majority or minority of the population would say "yes." Report and interpret the P-value.

c. Construct and interpret a 99% confidence interval for the population proportion who would say "yes."

1.9 A sample of women suffering from excessive menstrual bleeding have been taking an analgesic designed to diminish the effects. A new analgesic is claimed to provide greater relief. After trying the new analgesic, 40 women reported greater relief with the standard analgesic, and 60 reported greater relief with the new one.

a. Test the hypothesis that the probability of greater relief with the standard analgesic is the same as the probability of greater relief with the new analgesic. Report and interpret the P-value for the two-sided alternative. (Hint: Express the hypotheses in terms of a single parameter. A test to compare matched-pairs responses in terms of which is better is called a *sign test*.)

b. Construct and interpret a 95% confidence interval for the probability of greater relief with the new analgesic.

1.10 Refer to the previous exercise. The researchers wanted a sufficiently large sample to be able to estimate the probability of preferring the new analgesic to within 0.08, with confidence 0.95. If the true probability is 0.75, how large a sample is needed to achieve this accuracy? (Hint: For how large an n does a 95% confidence interval have margin of error equal to about 0.08?)

1.11 When a recent General Social Survey asked 1158 American adults, "Do you believe in Heaven?", the proportion who answered yes was 0.86. Treating this as a random sample, conduct statistical inference about the true proportion of American adults believing in heaven. Summarize your analysis and interpret the results in a short report of about 200 words.

1.12 To collect data in an introductory statistics course, recently I gave the students a questionnaire. One question asked whether the student was a vegetarian. Of 25 students, 0 answered "yes." They were not a random sample, but let us use these data to illustrate inference for a proportion. (You may wish to refer to Section 1.4.1 on methods of inference.) Let π denote the population proportion who would say "yes." Consider H_0: $\pi = 0.50$ and H_a: $\pi \neq 0.50$.

a. What happens when you try to conduct the "Wald test," for which $z = (p - \pi_0)/\sqrt{[p(1-p)/n]}$ uses the *estimated* standard error?

b. Find the 95% "Wald confidence interval" (1.3) for π. Is it believable? (When the observation falls at the boundary of the sample space, often Wald methods do not provide sensible answers.)

c. Conduct the "score test," for which $z = (p - \pi_0)/\sqrt{[\pi_0(1-\pi_0)/n]}$ uses the *null* standard error. Report the P-value.

d. Verify that the 95% score confidence interval (i.e., the set of π_0 for which $|z| < 1.96$ in the score test) equals (0.0, 0.133). (Hint: What do the z test statistic and P-value equal when you test H_0: $\pi = 0.133$ against H_a: $\pi \neq 0.133$.)

1.13 Refer to the previous exercise, with $y = 0$ in $n = 25$ trials.

a. Show that ℓ_0, the maximized likelihood under H_0, equals $(1 - \pi_0)^{25}$, which is $(0.50)^{25}$ for H_0: $\pi = 0.50$.

b. Show that ℓ_1, the maximum of the likelihood function over all possible π values, equals 1.0. (Hint: This is the value at the ML estimate value of 0.0.)

c. For H_0: $\pi = 0.50$, show that the likelihood-ratio test statistic, $-2\log(\ell_0/\ell_1)$, equals 34.7. Report the P-value.

d. The 95% likelihood-ratio confidence interval for π is (0.000, 0.074). Verify that 0.074 is the correct upper bound by showing that the likelihood-ratio test of H_0: $\pi = 0.074$ against H_a: $\pi \neq 0.074$ has chi-squared test statistic equal to 3.84 and P-value $= 0.05$.

1.14 Sections 1.4.4 and 1.4.5 found binomial P-values for a clinical trial with $y = 9$ successes in 10 trials. Suppose instead $y = 8$. Using the binomial distribution shown in Table 1.2:

a. Find the P-value for (i) H_a: $\pi > 0.50$, (ii) H_a: $\pi < 0.50$.

b. Find the mid P-value for (i) H_a: $\pi > 0.50$, (ii) H_a: $\pi < 0.50$.

c. Why is the sum of the one-sided P-values greater than 1.0 for the ordinary P-value but equal to 1.0 for the mid P-value?

1.15 If Y is a variate and c is a positive constant, then the standard deviation of the distribution of cY equals $c\sigma(Y)$. Suppose Y is a binomial variate, and let $p = Y/n$.

a. Based on the binomial standard deviation for Y, show that $\sigma(p) = \sqrt{[\pi(1-\pi)/n]}$.

b. Explain why it is easier to estimate π precisely when it is near 0 or 1 than when it is near 0.50.

1.16 Using calculus, it is easier to derive the maximum of the log of the likelihood function, $L = \log \ell$, than the likelihood function ℓ itself. Both functions have maximum at the same value, so it is sufficient to do either.

a. Calculate the log likelihood function $L(\pi)$ for the binomial distribution (1.1).

b. One can usually determine the point at which the maximum of a log likelihood L occurs by solving the *likelihood equation*. This is the equation resulting from differentiating L with respect to the parameter, and setting the derivative equal to zero. Find the likelihood equation for the binomial distribution, and solve it to show that the ML estimate equals $p = y/n$.

1.17 Suppose a researcher routinely conducts significance tests by rejecting H_0 if the P-value satisfies $P \leq 0.05$. Suppose a test using a test statistic T and right-tail probability for the P-value has null distribution $P(T = 0) = 0.30$, $P(T = 3) = 0.62$, and $P(T = 9) = 0.08$.

a. Show that with the usual P-value, the actual probability of type I error is 0 rather than 0.05.

b. Show that with the mid P-value, the actual probability of type I error equals 0.08.

c. Repeat (**a**) and (**b**) using $P(T = 0) = 0.30$, $P(T = 3) = 0.66$, and $P(T = 9) = 0.04$. Note that the test with mid P-value can be "conservative" [having actual P(type I error) below the desired value] or "liberal" [having actual P(type I error) above the desired value]. The test with the ordinary P-value cannot be liberal.

1.18 For a given sample proportion p, show that a value π_0 for which the test statistic $z = (p - \pi_0)/\sqrt{[\pi_0(1 - \pi_0)/n]}$ takes some fixed value z_0 (such as 1.96) is a solution to the equation $(1 + z_0^2/n)\pi_0^2 + (-2p - z_0^2/n)\pi_0 + p^2 = 0$. Hence, using the formula $x = [-b \pm \sqrt{(b^2 - 4ac)}]/2a$ for solving the quadratic equation $ax^2 + bx + c = 0$, obtain the limits for the 95% confidence interval in Section 1.3.4 for the probability of success when a clinical trial has nine successes in 10 trials.